Analysis of Airline Delay and Cancellation Data, 2009 - 2018

Paul Smith

Western Governors University

**Table of Contents**

[A. Project Overview 3](#_Toc114439248)

[A1. Research Question or Organizational Need 3](#_Toc114439249)

[A2. Context and Background 3](#_Toc114439250)

[A3. Summary of Published Works 3](#_Toc114439251)

[A3a. Relation of Published Works to Project 3](#_Toc114439252)

[A4. Summary of Data Analytics Solution 3](#_Toc114439253)

[A5. Benefit to Organization and Decision-Making Process 3](#_Toc114439254)

[B. Data Analytics Plan 4](#_Toc114439255)

[B1. Goals, Objectives, and Deliverables 4](#_Toc114439256)

[B2. Scope of Project 4](#_Toc114439257)

[B3. Standard Methodology 4](#_Toc114439258)

[B4. Timeline and Milestones 4](#_Toc114439259)

[B5. Resources and Costs 4](#_Toc114439260)

[B6. Criteria for Success 4](#_Toc114439261)

[C. Design of Data Analytics Solution 6](#_Toc114439262)

[C1. Hypothesis 6](#_Toc114439263)

[C2. Analytical Method 6](#_Toc114439264)

[C2a. Justification of Analytical Method 6](#_Toc114439265)

[C3. Tools and Environments of Solution 6](#_Toc114439266)

[C4. Methods and Metrics to Evaluate Statistical Significance 6](#_Toc114439267)

[C4a. Justification of Methods and Metrics 6](#_Toc114439268)

[C5. Practical Significance 7](#_Toc114439269)

[C6. Visual Communication 7](#_Toc114439270)

[D. Description of Datasets 8](#_Toc114439271)

[D1. Source of Data 8](#_Toc114439272)

[D2. Appropriateness of Dataset 8](#_Toc114439273)

[D3. Data Collection Methods 8](#_Toc114439274)

[D4. Data Quality 8](#_Toc114439275)

[D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance 8](#_Toc114439276)

[D5a. Precautions 8](#_Toc114439277)

[E. Sources 9](#_Toc114439278)

# Project Overview

# A1. Research Question or Organizational Need

This project will analyze airline delay and cancellation data from 2009 to 2018 and determine the causes of flight delays. Flight delays cost the US economy a great deal of money ($32.9 billion in 2007) and determining the cause of these delays could allow this cost to be mitigated or reduced.

# A2. Context and Background

A 2010 report sponsored by the Federal Aviation Administration (FAA) analyzed a variety of cost components caused by flight delays. This included the cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The report concluded that the total cost of all US air transportation delays in 2007 was $32.9 billion. Clearly, flight delays are a serious and widespread problem in the US.

An exploration of multi-year (2009 – 2018) airline delay and cancellation data will be made to determine the causes of airline delays.

# A3. Summary of Published Works

In this section summarize the content of the 3 articles related to the project.

Various factors contribute to aircraft delays. One of these factors is the impact that weather, especially extreme weather, can have in delaying flights. In January 2019, an article titled, “Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States,” written by Christopher J. Goodman and Jennifer D. Small Griswold detailed the average impact that certain weather phenomena have on aircraft delays. They found that it is important to assess airports individually as their operations efficiencies are impacted due to differences in weather and airline operations. Their findings included:

1. Differences in airport weather and airline operations impact the efficiencies of airport operations and it is important to assess airports individually.
2. Understanding the differences in airport weather climatologies allows for an understanding of how inclement weather reduces efficiency.
3. Weather impacts delays and cancellations in a way consistent with climatological weather patterns.

Nicholas G. Rupp writes, in article titled, “Further Investigations into the Causes of Flight Delays,” that flight delays should be investigated from both the airline and passenger perspectives. Three conclusions are reported:

1. There are considerable differences when using alternative measures of flight delays – excess travel time which is the airline perspective and minutes of arrival/departure delay which is the passenger perspective.
2. Seating capacity, load factor, departure time, and distance have significant effects on flight delays.
3. After controlling for airport-specific effects, most estimations indicate that airport concentration at origination has longer departure and arrival delays.

Paul Blackwood wrote, in an article titled, “Understanding Flight Delays at U.S. Airports in 2010, Using Chicago O’Hare International Airport as a Case Study,” that flight delays negatively impact the environment, the economy, and society. After studying various delays at the named airport, he concluded:

1. Late aircraft accounted for 40% of all delays.
2. Carrier related issues accounted for 29% of all delays.
3. National airspace bottlenecks accounted for 25% of all delays.
4. Weather accounted for 6% of all delays.
5. Security issues accounted for 1% of all delays.

# A3a. Relation of Published Works to Project

In this sub-section, discuss how each published work relates to the project and its planning.

# A4. Summary of Data Analytics Solution

Restate the business problem/need and the extent to which the proposed project will address it. The connection between the problem and solution should be clear.

# A5. Benefit to Organization and Decision-Making Process

An organization does not exist in a vacuum. Who are the stakeholders/users/decision-makers? How will the project’s process (execution) and outcomes help them?

# Data Analytics Plan

# B1. Goals, Objectives, and Deliverables

State the project’s goal or goals. Each goal must have at least 2 objectives to mark the process for achieving the goal, and each objective must terminate in at least one deliverable. A single goal is OK. Feel free to include a simple table to illustrate the relationship.

# B2. Scope of Project

Clearly state what your solution includes, and what it does not include.

# B3. Standard Methodology

Discuss your project management methodology and how you plan to proceed in each of its phases. Example methodologies are Waterfall, ADDIE, SDLC, CRISP-DM, SEMMA and KDD.

# B4. Timeline and Milestones

Present a table showing for each milestone its projected start and end dates, and its projected duration:

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Projected Start Date** | **Projected End Date** | **Duration (days/hours)** |
|  |  |  |  |
|  |  |  |  |

# B5. Resources and Costs

Present a table showing personnel, technology, and infrastructure costs. Add any needed descriptive detail below the table.

# B6. Criteria for Success

How will you measure the success and effectiveness of outcomes? Present at least 3 criteria for success, including how you will collect the data for each criterion and what constitutes success. Summarize your discussion in a table:

|  |  |  |
| --- | --- | --- |
| **Criterion/Metric** | **Required Data** | **Cut Score for Success** |
|  |  |  |
|  |  |  |
|  |  |  |

# Design of Data Analytics Solution

# C1. Hypothesis

Clearly state the hypothesis/need identified in A1.

# C2. Analytical Method

Clearly identify the descriptive or predictive method to be used.

Examples of acceptable descriptive methods:

* K-means clustering
* Hierarchical clustering
* Other clustering methods (max distance, min distance, etc.)
* PCA (all variables need to be numeric)
* MCA (all variables need to be factorized)
* FAMD
* Logistic regression with interpretations of estimated coefficients
* Multi-linear regression with interpretations of estimated coefficients

Examples of appropriate non-descriptive methods:

* Logistic regression
* Decision trees
* Random forest
* Neural network
* Multi-linear regression (if they provide a strong justification – like ease of interpretability)

# C2a. Justification of Analytical Method

Discuss why you chose this method. What makes it appropriate for this data and the question it is intended to address.

# C3. Tools and Environments of Solution

Identify the tool to be used for data extraction and why it is selected over other tools.

# C4. Methods and Metrics to Evaluate Statistical Significance

What are the methods and metrics you will use in the tool, and how is statistical significance determined?

# C4a. Justification of Methods and Metrics

Justify the appropriateness of these metrics and methods as the correct ones to use.

# C5. Practical Significance

Statistical significance indicates the existence of an effect, while practical significance deals with the magnitude of the effect. How practical is this result in the context of decision-making? Present examples of how this result will matter in practical use.

# C6. Visual Communication

What graphical representations will you use to communicate the findings? Why use these over others that may be chosen?

# Description of Datasets

# D1. Source of Data

The source of my data is [www.kaggle.com](http://www.kaggle.com). This website provides many datasets.

List the dataset(s) used for analysis. List them in **E. Sources** and cite them here, all in APA format.

# D2. Appropriateness of Dataset

These datasets are appropriate for addressing the project goal since they contain the proper fields for analysis. Using these datasets is the most appropriate way to provide my analysis with the data it will need while not relying on propriety data, since it is derived from data published by the US Government.

# D3. Data Collection Methods

Offer a discussion of how data was collected, including the steps taken to insure accuracy and quality. What are the advantages and disadvantages of collecting data using this methodology? Also discuss any challenges in the collection process and how they were overcome.

# D4. Data Quality

Discuss issues with the quality of the data and the need for interventions. Missing data, outliers, formatting issues, dirty data, and other data quality categories could be discussed.

# D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance

These considerations could relate to the problem/need, data sensitivity and nature, the industry within which the problem/need exists, etc.

# D5a. Precautions

Discuss guidelines related to data preparation, analysis, storage, access, and dissemination.

# Sources

Rupp, Nicholas G. (2007, May 20). Further Investigations into the Causes of Flight Delays. Retrieved August 30, 2022, from <https://economics.ecu.edu/wp-content/pv-uploads/sites/165/2019/07/ecu0707.pdf>

Blackwood, Paul. (2012, April). Understanding Flight Delays at U.S. Airports in 2010, Using Chicago O’Hare International Airport as a Case Study. Retrieved August 20, 2022, from <https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1033&context=masters_theses>

Goodman, Christopher J. and Small Griswold, Jennifer D. (2019, January). Meteorological Impacts on Commercial Aviation Delays and Cancellations in the Continental United States. Retrieved August 30, 2022, from <https://journals.ametsoc.org/view/journals/apme/58/3/jamc-d-17-0277.1.xml?tab_body=pdf>

Milan, Tomin. (2020). Airline Delay and Cancellation Data 2018. [www.kaggle.com](http://www.kaggle.com). https://www.kaggle.com/code/milantomin/airline-delay-and-cancellation-data-

2018/data?select=2009.csv